EFFICIENT EXCITATION MODEL AND FAST SELECTION
IN CELP CODING OF SPEECH
M. DELPRAT, M. LEVER and C. GRUET
MATRA Communication
Rue J.P. Timbaud, B.P. 26
78392 Bois d’Arcy Cedex
FRANCE

ABSTRACT
The paper discusses several new approaches for efficiently modeling and selecting the excitation in CELP coding of speech. Modified error criteria and structured codebooks lead to a wide range of complexity reduction methods, that are evaluated in terms of quality and computational requirements. A very low complexity, though high quality, Regular Pulse (RP) CELP technique is then derived. Finally we address the design of a robust 6 kbps RPCELP coder for mobile radio communications.

I. INTRODUCTION
Low bit rate coding techniques are currently an important topic in speech research, because of the wide range of emerging applications such as narrow band digital speech transmission or voice messaging. In designing of a low bit rate speech coder, the main issue is to achieve good quality with low enough complexity to allow for real time implementation. Moreover, in mobile telephony, the coder must also be robust to adverse transmission conditions (background noise, transmission errors...). Some existing coders meet these requirements [1], except for output speech quality which must be improved for general public applications.

In the past years, much has been done to improve the quality of speech coders at low bit rate. In that field, Code-Excited Linear Prediction (CELP) is undoubtedly the most popular technique [2]. In CELP coding, the speech signal is modelled as a random process with a slowly varying power spectrum. Synthetic speech is produced by filtering successive innovation sequences through long and short term predictors (figure 1). For each block, the optimum innovation sequence is selected from a codebook of vectors using an analysis by synthesis procedure with a perceptual error criterion. While this technique ensures that the reconstructed speech is close (in the subjective sense) to the original one, it also results in huge complexity. More recently, several related schemes with a reduced computational load have been proposed [3,4,5]. However, efforts are still necessary to further reduce complexity and to improve basic CELP quality at low bit rate (below 8 kbps).

The paper describes several new approaches for modeling and selecting the excitation in CELP coders, with the aim of reducing complexity while maintaining high quality synthetic speech. First, several error minimization criteria are compared; the role of the "memory" of linear prediction (LP) filters is pointed out and a convenient perceptual weighting filter is introduced. Section III deals with the excitation model, presenting a new Regular Pulse (RP) innovation codebook and comparing several long term prediction configurations. Section IV describes a wide range of fast codebook search algorithms, relying on modified error criteria and/or codebook structure. Finally section V addresses the design of a robust and very low complexity 6 kbps Regular Pulse CELP coder.

II. ERROR MINIMIZATION CRITERION
II.1. Selection of the optimum innovation sequence.
In the original CELP [2], the optimum innovation sequence is selected by filtering each possible codeword $c_k$, scaled by a gain factor $G_k$, through both long and short term predictors. The resulting synthetic signal is compared with the original one and the difference signal is processed through the perceptual weighting filter $W(z)$:

$$ W(z) = A(z) / A(z/H) $$

(1)

with $\gamma$ around 0.8. The codeword that minimizes the weighted error signal energy is then selected for the current block. An equivalent structure for the codebook search [3,5] is represented in figure 2.

![Figure 1: speech synthesis in CELP coders.](image1)

![Figure 2: modified CELP analysis procedure.](image2)

The weighted error signal energy is now expressed as

$$ E(k) = \| x \|_2^2 - \| z_k \|_2^2 = \| x - G_k H c_k \|_2^2 $$

(2)

where $x$ represents the perceptually weighted original signal with the contribution from past excitations subtracted and $H$ is the impulse response matrix of the weighted synthetic filter $1/A(z/H)$. The optimum innovation sequence $(c_k, G_k)$ which minimizes $E(k)$ in equation (2) is then determined in two steps:

1) Find the index $k_0$ which maximizes the weighted inner product $P_w(k)$:

$$ P_w(k) = (x^H H c_k) / \| H c_k \|_2 $$

(3)

2) Compute the related gain $G_{k_0}$:

$$ G_{k_0} = P_w(k_0) / \| H c_{k_0} \|_2 $$

(4)
II.2. Covariance versus autocorrelation approach.

The minimization of $E(h)$ described above is traditionally carried out over the current block, of length $L$. It is referred to as the "covariance" method and in this case the impulse response matrix $H$ is a $LxL$ lower triangular Toeplitz matrix given by

$$H = \begin{bmatrix}
h(0) & 0 & 0 & \ldots & 0 \\
h(1) & h(0) & 0 & \ldots & 0 \\
h(2) & h(1) & h(0) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
h(L-1) & h(L-2) & h(L-3) & \ldots & h(0)
\end{bmatrix}$$

where $h(i)$ is the impulse response of $1/A(z^f)$. Another approach \[6,7,8\] is to consider longer sequences of residual and excitation signals, completing the $L$ actual samples with $J$ zeros, where $J$ is chosen in such a way that $h(i)$ is practically zero for $i > J$. In this "autocorrelation" approach, $H$ becomes a $(L+J)x(L+J)$ Toeplitz matrix:

$$H = \begin{bmatrix}
h(0) & 0 & 0 & \ldots & 0 \\
h(1) & h(0) & 0 & \ldots & 0 \\
h(2) & h(1) & h(0) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
h(J) & h(J-1) & h(J-2) & \ldots & h(0) \\
0 & h(J) & h(J-1) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & h(J)
\end{bmatrix}$$

The terminology used here for the error minimization approach refers to LPC analysis. In the autocorrelation method the covariance terms $R_{ij}$ of the impulse response $h$ involved in the error minimisation procedure become autocorrelation terms $R_{ii+j}$ (stationarity assumption), leading to interesting symmetry properties \[7\], that will be exploited in section IV for the design of fast search algorithms. In fact, the covariance method results in a more accurate matching at the beginning of the block than at the end, whereas in the autocorrelation method all the excitation samples are evenly weighted.

In the autocorrelation method, when a rectangular window is applied over the block \[12\], signals are supposed to be zero outside the minimization interval and then, strictly speaking, the memory terms of the LP filters should be discarded \[8\]. Indeed, in our experiments, we noticed that the outputs of $1/A(z^f)$ with a zero excitation are quite similar for the original and synthetic signals. Though, the minimization procedure can take these memory terms into account by applying an extended rectangular window that begins $J$ samples before the current block \[4\].

On the contrary, since the covariance method does not evaluate the influence of current excitation on future blocks, the memory terms of the weighted synthesis filter play an important role in the analysis process.

We have compared both approaches and, as shown in table 1, their performances are very similar. However, as expected, the difference is more obvious when the memory terms are neglected. Results reported in table 1 have been obtained with a block size $L=20$ and a 7 bits Regular Pulse codebook. SNR values are averaged over 50 seconds of clean speech from 4 speakers (2 males, 2 females). In informal listening tests, no degradation could be heard when suppressing the memory terms in the autocorrelation method.

<table>
<thead>
<tr>
<th></th>
<th>Covariance</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With memory</td>
<td>Without</td>
</tr>
<tr>
<td>SNR</td>
<td>12.1</td>
<td>11.2</td>
</tr>
<tr>
<td>SNR-Seg</td>
<td>11.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

II.3. Perceptual weighting filter.

Perceptual weighting is known to significantly improve the subjective quality of low bit rate coders but its precise influence is not completely understood. In particular, the widely used form of $W(z)$ given in equation (1) is not the only expression of the perceptual weighting filter that leads to high subjective quality. We have studied the perceptual weighting performed by a convenient filter given by

$$W(z) = A(z) / C(z^f)$$

where $1/C(z)$ is an average low order linear short term speech predictor. The weighted synthesis filter is then modified in $1/C(z^f)$ whose coefficients are time invariant, a property that will be exploited in section IV for complexity reduction.

Such a filter has already been used in Multi Pulse coding of speech \[8\] and has proved its remarkable ability to provide almost equivalent subjective results. When applied to a CELP coder \[6\], it produces very small (if any) audible distortion in spite of a relatively lower SNR (up to 2.5 dB). An interpretation of the high performances obtained with $W(z)$ is that perceptual weighting is essentially efficient for voiced sounds and for these segments $W(z)$ is close to $W(z)$ (since $1/C(z)$ has the form of a smooth low pass filter).

III. EXCITATION MODEL

III.1. Innovation codebook design.

Recent studies have shown that the choosing the innovation codebook is not critical for the quality of the coder. Comparable performances have been obtained with a wide range of codebooks such as random codebooks with different statistical distributions \[2\], sparse codebooks \[5\] and binary or ternary codebooks that may be derived from algebraic codes \[3,4,10\]. On the other hand, as will be shown in section IV, coder complexity may greatly benefit from a strong codebook structure.

Good results achieved with codebooks including pulse sequences suggest using excitation sequences of length $L$ that have a regular structure consisting in $q$ equidistant pulses separated by $D-1$ zeros. The first pulse (initial phase $p$) is at one of the locations $0$ to $D-1$. The Regular Pulse (RP-) codebook \[6\], populated in a stochastic or deterministic manner, may be constituted of $K$ independent sequences or of the $D$ possible shifts of a basic set of $K/D$ RP-sequences with initial phase zero. In the latter case, each codeword $c_k$ is expressed as

$$c_k = \Delta_p d_m$$

where $\Delta_p$ is a $Lxq$ decimation matrix, function of the initial phase $p$ and $d_m$ is a $q$-dimensional vector with $k = p.(K/D) + m$.

A "binary" RP-codebook, built from the 2$L$ binary words of length $q$ (0 becoming -1), is particularly efficient to reduce both computational load and storing requirements (see section IV). Furthermore, RP-codebooks achieve comparatively high performances \[6\] because they allow for a better representation of the phase information in the excitation signal.

Regular Pulse excitation has already been used at higher rates in RPE coders \[8\] and such a low complexity coder has recently been chosen as a standard for the Pan-European mobile radio system \[11\]. Introducing a RP-codebook in a CELP coder can be seen as a RPE technique in which the pulse amplitudes are optimally vector quantized and in that sense may be considered as a Base Band CELP coding technique.
III.2. Long Term Prediction.

Long Term Prediction (LTP) is an essential feature of CELP coders since it produces most of the output signal energy [9]. LTP can be performed in many ways with a wide range of quality, complexity and bit rate.

Regarding LTP analysis, we found that the closed loop approach introduced in [13] outperforms by more than 1 dB on average the traditional open loop approach based on autocorrelation of residual signal. However the closed loop approach leads to a huge complexity and requires a high update rate of the parameters. When using a LTP frame length larger than the minimum value of the LTP delay, the strategies we tested to replace the unknown excitation samples (either by residual samples or by predicted excitation samples) did not prove to be very efficient. We also evaluated the open loop cross-correlation approach used in [11] and obtained slightly lower performances than with the autocorrelation approach, though the difference is less obvious with a full quantized coder. For all methods, we found it is essential to use normalized correlation terms when the LTP frame length is small.

Finally, we concluded that a simple LTP scheme (single tap, open loop analysis, relatively low update rate) is suitable for the design of a low bit rate CELP coder, both for complexity and quality purposes (since the available bit rate may be more efficiently devoted to the innovation codebook).

IV. FAST CODEBOOK SEARCH ALGORITHMS

A high complexity remains in the modified CELP structure described in section II due to the filtering of all codewords to find the one which maximizes the inner product $P_k(x)$ (eq. 3). Major ways to reduce this amount of computations are as follows: to exploit the structure of particular codebooks (e.g. algebraic [3] or simply binary [4, 10]), to choose a suitable perceptual weighting filter [6], to modify the error criterion [4, 6]. We present below different approaches that may be combined to speed up the search procedure in the codebook. A fast and efficient scheme called RPCELP is then derived.

IV.1. Time invariant weighted synthesis filter.

The complexity considerably decreases when the weighted synthesis filter is fixed, as proposed in section I.I. The filtered vectors $H.c_k$ and the weighting factors $P(k)$ involved in the inner product may all be precomputed and stored. A very simple but more memory consuming method (M1) is then obtained.

IV.2. Suitable error criterion (autocorrelation method).

In the autocorrelation approach defined in section II signal $x$ is expressed as

$$x = Hr$$

where $r$ represents the LP residual vector with the LTP contribution subtracted. The weighted inner product becomes:

$$P_k(x) = y.c_k / \|H.c_k\|^2$$

where $y = H.H.r$ can be obtained as the result of a particular filtering operation (with time varying coefficients in general case) of each residual vector $r$ before the codebook search [6]. This is illustrated in figure 3.

The remaining weighting term $\|H.c_k\|^2$ may be neglected in a first approximation. This leads to a very fast but suboptimal method (M2) in which the unnormalized inner product $P_k(x) = y.c_k$ is to be maximized. This approximation has a small impact on the perceived speech quality if the excitation sequences are such that all the $H.c_k$ have almost the same value. This is the case for the RP-codebook defined in I.I.

An optimum scheme requires the computation and storage of the norms of the filtered vectors once per frame [51], (method M3). One way to efficiently perform this filtering operation is to convolve the excitation sequences with the (truncated) impulse response $h$ of the time-varying weighted synthesis filter. This product may be obtained as the sum of the contributions of each excitation pulse. The number of operations varies with the number of pulses and this method (M4) is rather fast in the case (e.g. RP-codebook) of sparse codewords [10]. Moreover, the norms may be jointly computed using the fact that sequences with the same values up to sample $n$ lead to identical partial sums up to index $n$.

### Figure 3: Fast CELP structure with modified error criterion.

IV.3. The RPCELP approach.

When introducing a regular pulse excitation, together with the autocorrelation approach, the codebook structure can be exploited to further accelerate the search procedure. The normalization factors become:

$$\|H.c_k\|^2 = q_1.H.H.c_k = d_{m^2}R_D$$

where $R_D = \Delta^H_D.H.H.$$\Delta_{FD}$ is a $q_xq$ symmetrical Toeplitz matrix whose $i$th diagonal term is $R((i-1)D)$. Note that $R$ is independent of the phase $p$ as a result of the autocorrelation method defined above.

Moreover, $R_D$ can be forced to a diagonal matrix, using a reasonable approximation on the weighted synthesis filter: for instance, its impulse response can be shortened in order that $h(n) = 0$ for $n \geq D$ as described in [8]. With the normalization by $R(0)$, the matrix $R_D$ becomes the identity matrix and equation (11) gives

$$\|H.c_k\|^2 = \|d_m\|^2$$

Assuming the codewords are normalized, the search procedure comes down to maximize the inner product $P(k) = y.c_k$, which represents a small amount of computations (all the more because the codewords are sparse). This RPCELP method was first introduced in [6].


Besides, with a binary RP codebook the optimum codeword is efficiently determined in a two-steps procedure:

1) Find the phase $p$ which maximizes $M(p)$ with

$$M(p) = \sum_i y(p+iD), \text{ sum over } i=0, \ldots, q-1$$

2) Choose the vector $d_m$ such that $y.c_k = M(p)$:

$$d_m(i) = \text{sign of } y(p+iD), \text{ for } i = 0, \ldots, q-1$$

The related gain is then given by $G_k = M(p)d_m$, since $\|H.c_k\|^2 = q$ for any $k$.

The perceived speech quality produced by the fast RPCELP method has been found to be equivalent to that of the original CELP.
IV.5. Complexity evaluation.

We have evaluated the computational requirements of the codebook search for the different methods described above. Table 2 gives both approximated analytical expressions and estimated values for a particular configuration.

Note that the complexity of the fast RPELP algorithm is practically independent of the codebook size K. This scheme is especially suitable for real-time implementation since it requires less than 20 operations per sample for the entire excitation determination (filtering of the residual signal, optimum index and gain computation).

Table 2: Codebook search complexity comparisons (number of operations per sample)

<table>
<thead>
<tr>
<th>Codebook size</th>
<th>Number of non-zero samples</th>
<th>LP order</th>
<th>Number of blocks per frame</th>
<th>Decimation factor</th>
<th>Block length</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=128</td>
<td>q=5</td>
<td>M=8</td>
<td>B=8</td>
<td>D=4</td>
<td>L=20</td>
</tr>
<tr>
<td>2560</td>
<td>128</td>
<td>270</td>
<td>162</td>
<td>32</td>
<td>5</td>
</tr>
</tbody>
</table>

The robustness of the coder to transmission errors has been evaluated and the specific sensitivity of each parameter has been tested. We found that protecting the most significant bits of the first LP filter coefficients and of the energy term (maximum gain) is sufficient to preserve good intelligibility, even with a high error rate (more than $10^{-3}$), but it does not maintain a high quality output speech. As a matter of fact, errors on LP parameters or on excitation parameters result in uniform degradation that gradually increases with error rate. However, it should be noted that the binary RP-codeword is especially robust against transmission errors since one error on a given bit of a codeword index (except for the phase information) produces only a single wrong pulse in the innovation sequence [4,6].

CONCLUSION

A class of speech CELP coding methods, based on a regular pulse excitation model, has been presented. We focused on the perceptual error criterion to derive several low complexity methods. Even the least complex of them, called fast RPELP, which has been fully quantized at 6 Kbps and implemented in 16 bits fixed point arithmetic, produces a reconstructed speech quality which is very close to that of the original CELP scheme. Therefore, it appears to be an efficient solution for low cost speech coding applications.

REFERENCES